Effect of context, rebinding and noise, on audiovisual speech fusion

Ganesh AttigoduChandrashekara,FrédéricBerthommier, OlhaNahorna, Jean Luc Schwartz,

GIPSA-Lab – DPC, UMR 5216 –CNRS University of Grenoble, France
ganesh.attigodu, olha.nahorna, Jean-Luc.Schwartz, Frederic.Berthommier@gipsa-lab.grenoble-inp.fr
http://www.gipsa-lab.grenoble-inp.fr

Abstract

In a previous set of experiments we showed that audio-visual fusion during theMcGurk effect may be modulated by context. Ashort context (2 to 4 syllables) composed of incoherent auditory and visual material significantly decreases theMcGurk effect. We interpreted this as showing the existence of an audiovisual "binding" stagecontrolling the fusionprocess, and we also showed the existence of a"rebinding" process when an incoherent material isfollowed by a short coherent material. In this work we evaluate the role of acoustic noise superimposed to the context and to the rebinding material. We use either a coherent or incoherent context, followed, if incoherent, by a variable amount of coherent "rebinding" material, with two conditions, either silent or with superimposed speech-shaped noise. The McGurktarget is presented with no acoustic noise. We confirm the existence of unbinding (lower McGurk effect with incoherent context) and rebinding (the McGurk effect is recovered with coherent rebinding). Noiseuniformly increases the rate of McGurk responses compared to the silentcondition. We conclude on the role of audiovisual coherence and noise in the binding process, in the framework of audiovisual speech scene analysis and the cocktail party effect.

Index Terms: audiovisual speech perception, McGurk effect, unbinding, rebinding, perception in noise

1. Introduction

It is known since long that the human brain combines visual and auditory information to better understand spoken language, particularly in the case of perception in noise [1-4]. A classical paradigm to demonstrate audiovisual fusion is provided by the "McGurk effect" in which a conflicting visual input modifies the perception of an auditory input, e.g. visual /ga/ added on auditory /ba/ leading to the percept of /da/ [5].

Audiovisual fusion in speech perception has long been considered as automatic [6, 7]. However a number of recent experiments have provided evidence that it is in fact under the control of attention in a broad sense, consideringthat various cognitive variables can modulate audiovisual integration [8-13].

1.1 Binding andunbinding in audiovisual fusion

While evidence for the non-automaticity of the fusion mechanism stays compatible with a one-stage architecture, some data suggest that audiovisual interactions could intervene at various stages in the speech decoding process [14-16]. Actually, audiovisual fusion could be conceived as a two-stage process, beginning by *binding* together the appropriate pieces of audio and video information, followed by integration per se [17]. The binding stage would occur early in the audiovisual speech processing chain enabling the listener to extract and group together the adequate cues in the auditory and visual streams, exploiting coherence in the dynamics of the sound and sight of the speech input.

To demonstrate the existence of this "binding" process we defined an experimental paradigm possibly leading to "unbinding". In thisparadigm(Figure 1) incongruent "McGurk" (A/ba/ + V/ga/) or congruent "ba" (A/ba/ + V/ba/) targets were preceded by coherent or incoherent audiovisual contexts [18]. The experimental results showed that the McGurkeffect (displaying the role of the visual input on phonetic decision)depends on the previous audiovisual context. Indeed, various kinds of incoherent contexts,such asacoustic syllables dubbed on video sentences, or phonetic or temporal modifications of the acoustic content of a regular sequence of audiovisual syllables, can significantly reduce the McGurk effect.Short incoherent context durations (even 1-syllable long) were sufficient to produce a significant amount of unbinding.On the contrary, coherent contexts let the McGurk effect stable, which suggests that there is possibly a "default mode" in which binding occurs (and hence produces the McGurk effect in isolation).

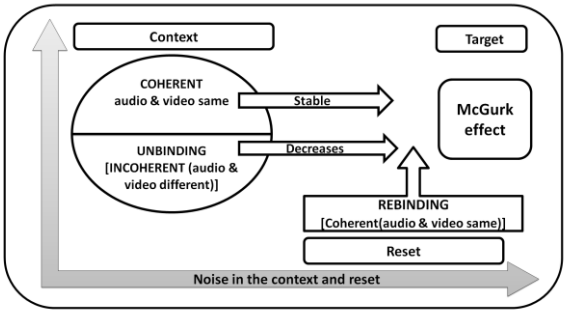


Figure 1: Experimental paradigm for displaying unbinding or rebinding mechanisms modulating the McGurk effect

1.2 Evidence for a rebinding process

Then we explored possible"rebinding" effects, searching forparadigmsable to reset the system and put it back in its default mode in which the McGurk effect would recover from unbinding. For this aim we prepared stimuli in which an incoherent context was followed by variable durations of a "reset" stimulus, either acoustic silence dubbed on fixed image, or coherent audiovisual material before the McGurktarget (Fig. 1). The results [19] showed thatthe

"silence + fixed image" reset did not provide any rebinding (the decrease in McGurk effect was not removed by reset). On the contrary, the coherent reset did produce rebinding, that is a significant increase in theMcGurkeffect, coming back to its "default" state for a coherence period of 4 syllables. A first objective of the present study is to confirm the existence of "unbinding-by-incoherence" and "rebinding-by-coherence" processes on the McGurk effect.

1.3 The role of noise in audiovisual fusion

The role of visual speech is particularly important in noise [1-4]. Noise also seems to modulate decision in the case of incongruent stimuli. Indeed, the McGurkeffectdecreases when the extraneous noise is visual, whereas it increases when the noise is auditory[20-24]. In the well-known "Fuzzy-Logical Model of Perception" (FLMP [6]) this is interpreted as due to the increasing ambiguity of the noisy component, which would automatically decrease its role in the fusion process. However, it could also be envisioned that there is a specific weighting process in which a given modality would be positively or negatively modulated in the fusion process depending on the noise in this modality [25, 26].In the first case fusion would only depend on stimuli while on the second case there would be in addition an evaluation of the conditions of perception resulting in a modification of the fusion process per se. Our reasoning here is that if noise is applied in the (context + reset) part of the stimulus in Fig. 1 but not on the target itself, if fusion only depends on stimuli, then the McGurk effect should not change since the McGurk target stays clear. If however fusion depends on a weighting process depending on the environment, then application of acoustic noise in the context part should result in increasing the role of vision in fusion, hence increasing the McGurk effect. The second objective of the present study is to test the role of noise on context, and its interaction with the binding/unbinding/rebinding processes.

2. Method

Globally, the experiment consisted in testing the McGurk effect in various kinds of contexts including: (i) a coherent vs. incoherent part to replicate unbinding – decrease of the McGurk effect – with incoherent contexts; (ii) in case of incoherent contexts, a coherent reset component to replicate rebinding – recovery of the McGurkeffect; (iii) addition of acoustic noise in one set of conditions, to test if noise added to the (context+reset) part could increase globally the McGurk effect.

2.1. Stimuli

The stimuli are described in Fig. 2. They are typically made (Fig. 2, top) of:

- an incoherent context (2 or 4 acoustic syllables superimposed on excerpts of video sentences matched for equal duration);
- followed by a reset stimulus consisting in 0, 1, 2 or 3 coherent audiovisual syllables;
- followed by a target which can be either a congruent audiovisual "ba" or a McGurkstimulus consisting in an audio "ba" dubbed on a video "ga".

A control stimulus, aimed at providing a reference for the McGurk effect, is provided by (Fig. 2, bottom):

- an coherent context (2 or 4 coherent audiovisual syllables;
- followed by a target which can be either a congruent audiovisual "ba" or a McGurkstimulus.

A series of audiovisual films were presented to participants in two blocks, one without acoustic noise and the other one with acoustic noise superimposed on all the context and reset parts of the stimuli. Noise consisted in speech-shaped noise0 dB SNR. The target parts always remained without noise.

Coherent context and reset material was constructed by pairing audiovisual syllables randomly selected within the following syllables ("pa", "ta", "va", "fa", "za", "sa", "ka", "ra", "la", "ja", "cha", "ma", "na"). In the incoherent context material, the auditory content was same, but the visual content was replaced by excerpts of video sentences matched in duration. The congruent "ba" target was used to ensure that participants were performing the speech task correctly and to serve as a baseline to contrast with the McGurk effect. The incongruent McGurktarget was produced by carefully synchronizing an auditory /ba/ with a video /ga/, precise temporal localization of the acoustic bursts of the original "ba" and "ga" stimuli providing the cue for synchonisation.

McGurktargetswere presented three times more than congruent "ba" targets, which served as controls. For each (context+reset) condition (2 context durations; 4 reset durations for incoherent context; 2 noise conditions; hence altogether 20 context conditions) there were 4 occurrences of a "ba" target and 12 occurrences of a McGurk target. Hence there were 320 sequences in total spread over 2 blocks of 10 min each.

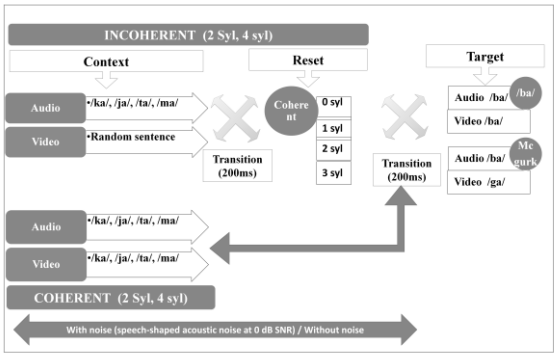


Figure 2.Description of the audiovisual material

2.2. Procedure

All experiments were carried out in a soundproof booth. Stimulus presentation and recording of responses were controlled by the Presentation software. The experiment consisted of two possible responses"ba" or "da" (with one button for "ba" and one for "da,") and the participants were instructed to constantly look at the screen and, each time a "ba" or a "da" was perceived, to immediately press the corresponding button. The films werepresented on a computer monitor with high-fidelity headphones set at a comfortable fixed level. The video stream was displayed at a rate of 25 images per second, the subject being positioned at about 50 cm from the screen. There were 5 different orders of the stimuli in the films, the order of the two blocks "silent" and "noise" was counterbalanced between subjects, and the response button was also interchanged between subjects.

2.3 Participants

Twenty participants (13 women and 7 men; 34 years) participated in the experiment. All of them were French native speakers, without any reported history of hearing disorders and with normal or corrected-to-normal vision. Written consent was obtained from each participant and all procedures were approved by the Grenoble Ethics Board (CERNI).

2.4Assumptions andanalyses

The experiment was focused on the role of context, reset and noise on the McGurk effect. For each (context, reset and noise) condition, each target and each subject, the amount of "ba" responses against "ba+da" responses was computed and used as an index of the subject's perception. Though response times were systematically recorded and processed, they will not be presented here.

We had three main assumptions, all involving McGurk stimuli (let us recall that "ba" targets are just there as controls.

- Firstly, incoherent context should produce unbinding and decrease the McGurk effect (hence increase the amount of "ba" responses) in respect to coherent context, whatever the context duration (2 or 4 syllables).
- Secondly, for incoherent context, coherent reset should produce rebinding and increase the McGurk effect (hence decrease the amount of "ba" responses), from 0 to 3 syllables of duration of the reset coherent stimulus.
- Thirdly, noise should enhance the role of vision and hence globally increase the McGurk effect (decrease the score of "ba" responses) whatever the context and reset.

3. Results

3.1 Global effect of target, noise and context duration in the incoherent context condition

On Fig. 3 we display the global scores (percentage of "ba" responses relative to "ba" + "da" responses) for all "ba" and McGurk targets, averaged over all incoherent context conditions (whatever context and reset duration), for silent vs. noise blocks. As expected (and as in previous experiments) the "ba" target leads to 100% "ba" responses. This is the case also in the coherent context. Therefore, for now on, we shall concentrate on McGurk targets.

The score is much lower for McGurk targets, with a score lower than 70% of "ba" responses (hence more than 30% "da" responses): this is the classical McGurk effect, which is known to produce such kinds of scores in French [27]. Globally, noise within context happens to decrease the "ba" scores and hence increase the McGurkeffect by around 15%.

We compared responses for McGurktargets depending on noise and on context duration (2 vs. 4 syllables). Even though the marginal difference was present in the context [F (1, 19)=7.17, P=0.014], there was no interaction effect between context and noise [F(1,19)=0.8, P=0.381]. This confirms [19] that the context duration has only little effect on the McGurkeffect, hence we shall average data for the two context durations in the next analysis.

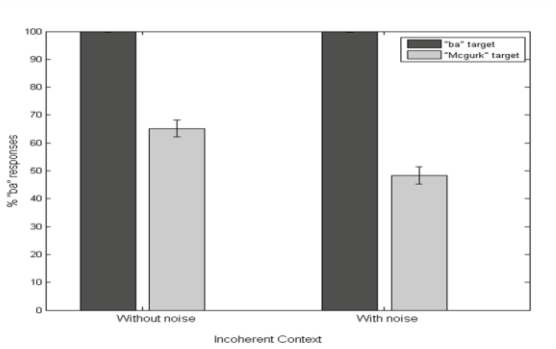


Figure 3. Percentage of "ba " responses for " ba " (in dark grey) and " McGurk " (in light grey) targets, in the "without noise" (left) vs. "with noise" (right) conditions for incoherent context.

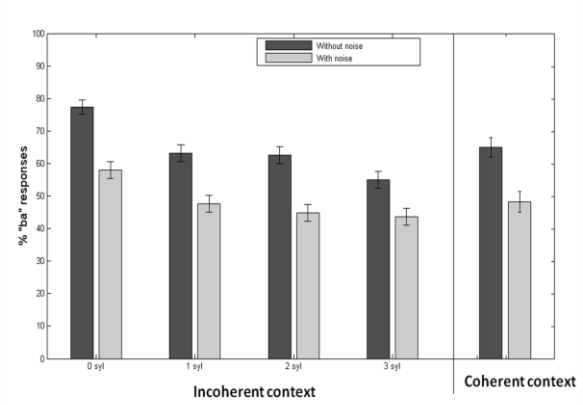
3.2 Assessing the effect of rebinding and noise onMcgurk targets

On Fig. 4 we display "ba" scores for McGurk targets in all conditions (averaging over the two context durations 2 and 4 syllables). Three major facts emerge from this figure.

- *Unbinding*: Focusing on the "without noise" condition (black bars), the score is higher (less McGurk effect) for incoherent context without reset (most left) than for coherent context (most right). This replicates previous findings in [18].
- *Rebinding*: Still focusing on the "without noise" condition (black bars), the score decreases (more McGurk effect) with increasing reset duration. For all reset durations higher than 1 the score has actually reached a value equal to or even lower than for coherent context. This replicates previous findings in [19].
- *Modulation by noise*: Comparing black (with noise) and grey (without noise) bars, noise decreases scores (increases the McGurk effect) for all conditions.

To assess the effects of rebinding and noise, we performed an analysis of variance with the factors subject (random-effect), reset (with noise & without noise) and reset duration (0, 1, 2 & 3 syllables). An arc(sin(sqrt)) transformation was applied on "ba" scores to ensure Gaussianity of the dependent variable.Subject, reset and reset duration factors are both statistically significant [subject: $F(19,18)=13.46$, $P<0.001$; reset: $F(1,19)=6.12$, $P<0.05$]; reset duration [$F(3,57)=14.82$, $P<0.001$]. There was no significant interaction between any pair of factors.

Therefore this confirms our previous results in unbinding and rebinding [18, 19]. It also shows that noise applied in the context part modifies the results of audiovisual fusion, with a global and more or less stable effect leading to an increase of about 15% in the McGurk effect whatever the context.



ts, in the "without noise" vs. "with noise" conditions for incoherent context with the four reset durations, compared with coherent context. The ANOVAwas performed only for the four reset durations and the two noise levels in the incoherent context.

4. Discussion

4.1 Unbinding, rebinding and noise in the audiovisual fusion process

This experiment enabled us confirm that context may modify the McGurk effect, through a process that we described by general binding principles, with "unbinding" with incoherent context and "rebinding" with coherent reset [17, 18, 19].

In the present study, the rebinding process was evaluatedwithout noise vs.with noise applied on the "context+reset" portions of the stimuli, while McGurk targets were systematically silent. It appears that noise systematically increases the Mcgurkeffect. Toour knowledge this is the first time such a result is obtained. This strongly suggests that noise in the McGurk effect, already displayed with noise applied on the target itself [20-24], intervenes not only at the level of the stimuli, but also at the level of the fusion process itself.

At this level, it is possible to come back to the models of audiovisual fusion available in the literature. Classical models consider that phonetic decision operates at a given representational stage and produces an integrated percept combining auditory and visual cues in a given way, possibly mediated by general attentional mechanisms. Our data on the binding process led us suggest that an additional computational stage should be incorporated before decision operates, involving online computation of some assessment of the coherence/ incoherence of the auditory and visual inputs, resulting in a "two-stage model" of audiovisual speech perception [17] (see Fig. 5).

The present results first add some information about the way coherence could be computed, involving a dynamics made of unbinding and rebinding stages with short constant times: indeed, less than one second of incoherent (2 syllables or less) suffice to produce unbinding, and less than one second on coherence (2 syllables or less) suffice to produce complete rebinding.

Furthermore, the results about noise suggest that noise, and probably more generally knowledge about the conditions of communication, also participate to the decision process by providing an enhancement of "efficient" modalities, not contaminated by noise, versus modalities where noise could contaminate the decision process (Fig. 5).

The present data suggest that the role of unbinding/rebinding on one hand, and noise-based selective weighting of each modality on the other hand, could play additional independent roles, according to the lack of interaction between noise and reset in Section 3.2. This will have to be confirmed in future experiments specifically dealing with this question.

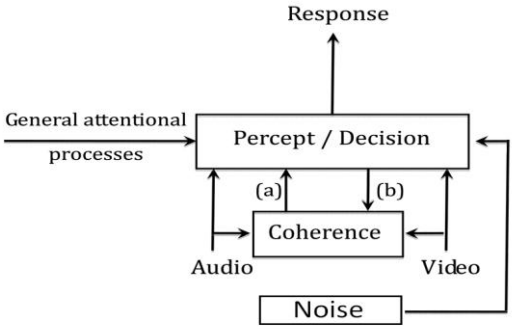


Figure 5: A two-stage model of audiovisual speech perception

4.2Future experiments

A number of further experimentswill have to extend the present data in various directions, involving e.g. more about the dynamics of unbinding and rebinding. Various proposals could also deal with reset mechanisms (such as changing speaker or the global communication setting), or specificity of the binding mechanism (could non-speech incoherent audiovisual material also produce unbinding?). The role of noise could also be further assessed by usingvisual noise. Indeed, some studies [10, 24] have manipulated the size or position of the face and found influence on the McGurk effect, showing in both that visual noise maydecrease the McGurkeffect just as auditory noise increases it. If our conjecture on the role of noise in Fig. 5 is correct, thus effect should also occur for visual noise added on the contextual part of the stimuli in the present paradigm.

Another very important extension concerns intelligibility in noise. The present paradigm was also an aim to progress towards the next important question that is to know if unbinding mechanisms would also decrease the beneficial effect of lipreading in noise. Future experiments will deal with targets consisting in ambiguous though coherent stimuli and test if an incoherent audiovisual context is able to remove the visual benefit. This will enable us incorporate the two-stage model inside a general question concerning the cocktail-party effect and what we propose to call "audiovisual speech scene analysis" [18].

5. Acknowledgements

This project has been supported by ANR Multistap (MULTISTability and binding in Audition and sPeech: ANR-08-BLAN-0167 MULTISTAP) and by Academic Research Community "Quality of life and ageing" (ARC 2) of the Rhône-Alpes Region, which provided a doctoral funding for Ganesh AttigoduChandrsheakra.

6. References

[1] Binnie, C. A., Montgomery, A. A., and Jackson, P. L., "Auditory and visual contributions to the perception of consonants", J. Speech Hear. Res., 17:619–630,1974

[2] Erber, N. P., "Auditory and audiovisual reception of words in low-frequency noise by children with normal hearing and by children with impaired hearing", J. Speech Hear. Res., 14:496–512, 1971.

[3] Grant, K. W., and Seitz, P., "The use of visible speech cues for improving auditory detection of spoken sentences", J. Acoust. Soc. Am., 108:1197–1208, 2000.

[4] Sumby, W. H, and Pollack, I., "Visual contribution to speech intelligibility in noise", J. Acoust. Soc. Am., 26: 212–215, 1954.

[5] McGurk, H., and MacDonald, J., "Hearing lips and seeing voices", Nature., 265:746–748, 1976.

[6] Massaro, D., "Speech perception by ear and eye", Hillsdale: LEA, 1987

[7] Soto-Faraco, S., Navarra, J., and Alsius, A., "Assessing automaticity in audiovisual speech integration: evidence from the speeded classification task", Cognition., 92; B13–B23, 2004.

[8] Alsius, A et al., "Audiovisual integration of speech falters under high attention demands", Curr. Biol., 15:839–843, 2005.

[9] Buchan, J.N. and Munhall, K.G., "The Effect of a Concurrent Cognitive Load Task and Temporal Offsets on the Integration of Auditory and Visual Speech Information", Seeing Perceiving., 25(1):87-106, 2012.

[10] Colin, C., Radeau-Loicq, M. and Deltenre, P., "Top-down and bottom-up modulation of audiovisual integration in speech", Eur. J. Cogn. Psychol., 17(4):541-560, 2005.

[11] Mozolic, J.L et al., "Modality-specific selective attention attenuates multisensory integration", Exp. Brain. Res., 184(1):39-52, 2008.

[12] Navarra, J et al., "Assessing the role of attention in the audiovisual integration of speech", Informational Fusion., 11(1):4-11, 2009

[13] Tiippana, K., and Andersen, T. S. (2004). Visual attention modulates audiovisual speech perception. Eur. J. Cogn. Psychol., 16:457-472, 2004.

[14] Bernstein, L., Auer, E., and Moore, J., "Audiovisual speech binding: convergence or association?" In G.A. Calvert et al(Eds) The handbook of multisensory processes, Cambridge: The MIT Press, 203–224, 2004;

[15] Grant, K. W., and Seitz, P., "The use of visible speech cues for improving auditory detection of spoken sentences", J. Acoust. Soc. Am., 108:1197–1208, 2000

[16] Van Wassenhove, V., Grant, K., and Poeppel, D., "Visual speech speeds up the neural processing of auditory speech", PNAS., 102:1181–1186.

[17] Berthommier, F. "A phonetically neutral model of the low-level audiovisual interaction", Speech. Comm., 44:31–41, 2004.

[18] Nahorna, O., Berthommier, F., and Schwartz, J.L., "Binding and unbinding the auditory and visual streams in the McGurk effect", J. Acoust. Soc. Am., 132:1061-1077, 2012.

[19] Nahorna, O. "Audiovisual speech binding", PhD Thesis, Grenoble University (to appear),2013.

[20] Sekiyama, K., and Tohkura, Y., "McGurk effect in non-English listeners: Few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility", J. Acoust. Soc. Am;, 90:1797-1805, 1991.

[21] Sekiyama, K. and Tohkura, Y., "Inter-language differences in the influence of visual cues in speech perception", J. Phonetics., 21:427-444, 1993.

[22] Sekiyama, K., "Difference in auditory-visual speech perception between Japanese and Americans: McGurk effect as a function of incompatibility". J. Acoust. Soc. Jap., 15:143-158, 1994.

[23] Fixmer, E., and Hawkins, S., "The influence of quality of information on the McGurk effect", in Proceedings AVSP'1998., Terrigal, Australia, pp. 27-32.

[24] Kim, J., and Davis, C., "Audiovisual speech processing in visual speech noise", in Proceedings AVSP'2011., Volterra, Italy, pp. 73-76.

[25] Heckmann, M., Berthommier, F., and Kroschel, K., "Noise Adaptive Stream Weighting in Audio-Visual Speech Recognition", EURASIP J. Appl. Signal Processing., 11:1260–1273, 2002

[26] Huyse, A, Berthommier, F, and Leybaert, J., "Degradation of labial information modifies audiovisual speech perception in cochlear-implanted children",Ear. Hear., 34(1):110-21, 2012.

[27] Cathiard, M. A., Schwartz, J. L., &Abry, C. "Asking a naive question about the McGurk Effect: why does audio [b] give more [d] percepts with visual [g] than with visual [d]?", in Proceedings of AVSP-2001., pp.138–142.